the German tank problem

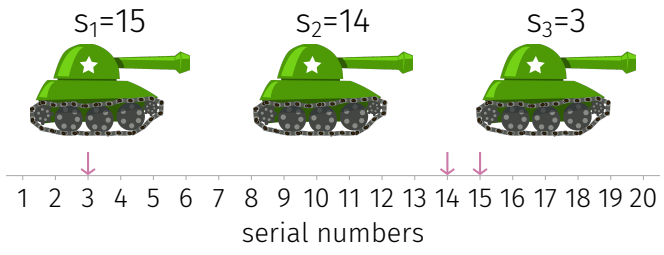
a Bayesian treatment

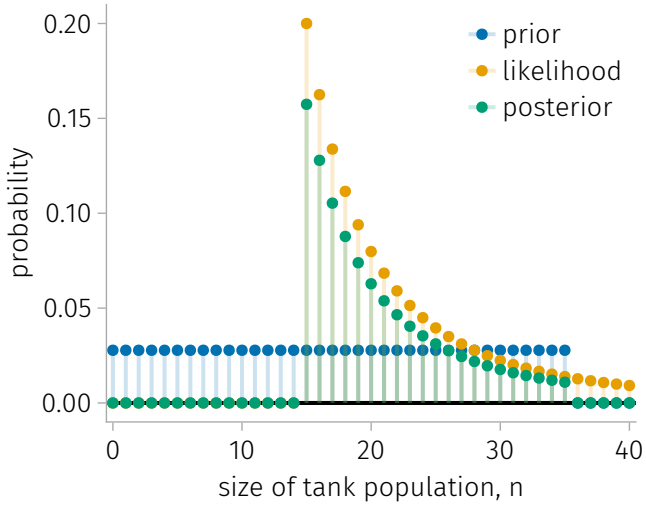
Cory M. Simon

Assistant Professor
School of Chemical, Biological, and Environmental Engineering
Oregon State University
simonensemble.github.io
@CoryMSimon

read the preprint!

C. Simon. ”A Bayesian treatment of the German tank problem.” *arXiv*. (2023)

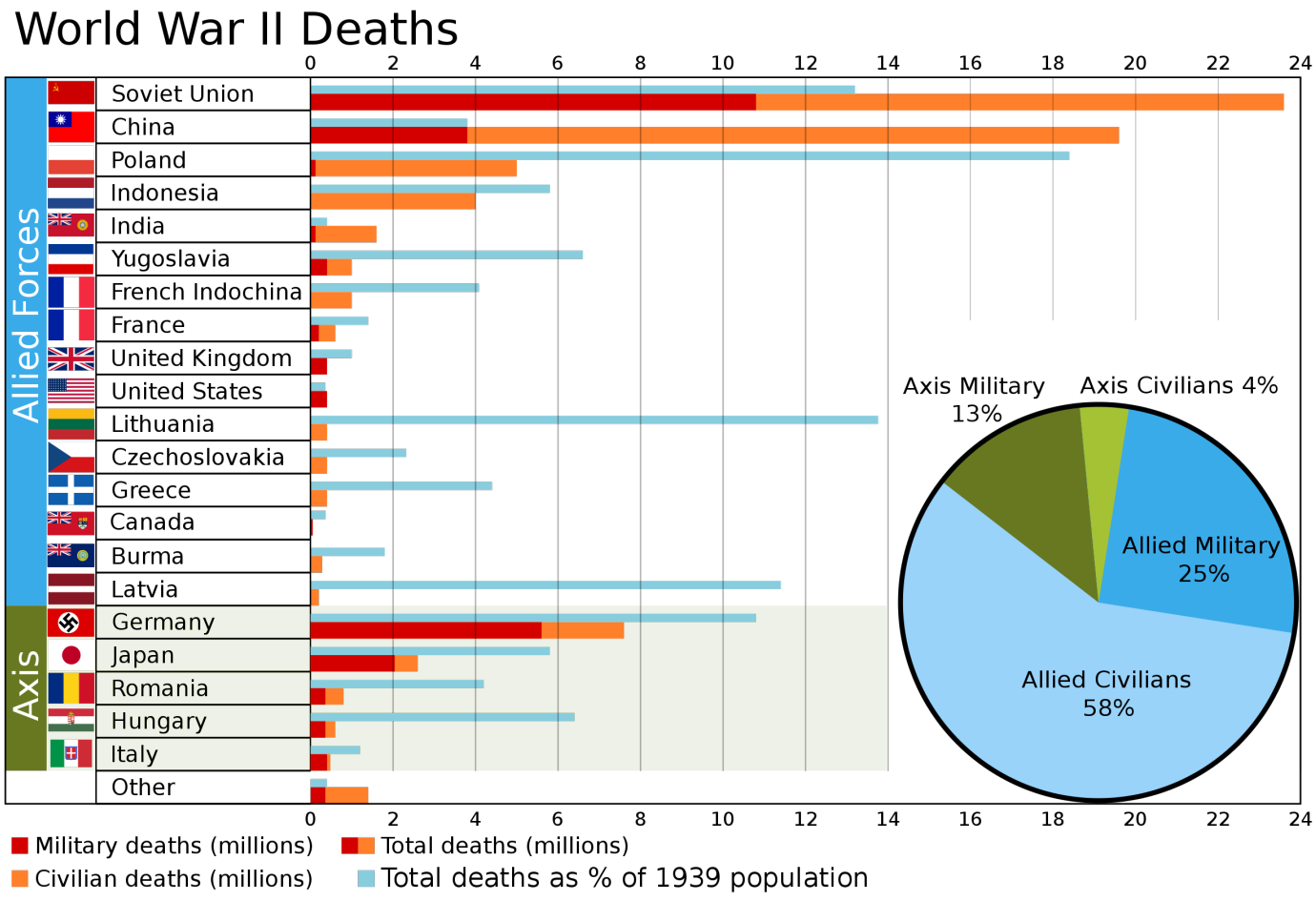




feedback welcome.

historical context

World War II (1939-1945)



source: Wikipedia

estimating Germany's armament production

context: the Allies wish to estimate Germany's production of tanks, tires, rockets, etc.

the German Panther tank. source: Wikipedia

[1] R. Ruggles and H. Brodie, "An empirical approach to economic intelligence in World War II", Journal of the American Statistical Association **42**, 72–91 (1947).

estimating Germany's armament production

context: the Allies wish to estimate Germany's production of tanks, tires, rockets, etc.

the German Panther tank. source: Wikipedia

conventional methods were unreliable and/or contradictory[1].

- extrapolating prewar manufacturing capabilities
- reports from secret sources
- interrogating prisoners of war

[1] R. Ruggles and H. Brodie, "An empirical approach to economic intelligence in World War II", Journal of the American Statistical Association **42**, 72–91 (1947).

Germany's practice of inscribing military equipment with serial numbers

Germany marked their military equipment with serial numbers and codes for the date and/or place of manufacture[2] to:

- facilitate handling spare parts
- trace defective equipment/parts back to the manufacturer for quality control.

[2] R. Ruggles and H. Brodie, "An empirical approach to economic intelligence in World War II", Journal of the American Statistical Association **42**, 72–91 (1947).

Germany's practice of inscribing military equipment with serial numbers

Germany marked their military equipment with serial numbers and codes for the date and/or place of manufacture[2] to:

- facilitate handling spare parts
- trace defective equipment/parts back to the manufacturer for quality control.

💡 these markings on a captured sample of German equipment conveyed information to British and American economic intelligence agencies about Germany's production of it.

[2] R. Ruggles and H. Brodie, "An empirical approach to economic intelligence in World War II", Journal of the American Statistical Association **42**, 72–91 (1947).

serial number analysis for German tank production

the Allies collected serial numbers on the chassis, engines, gearboxes, and bogie wheels of samples of tanks by inspecting captured tanks and examining captured records[3].

Monthly production of tanks by Germany.

	estimates
date	conventional intelligence
Jun. 1940	1000
Jun. 1941	1550
Aug. 1942	1550

[3] R. Ruggles and H. Brodie, "An empirical approach to economic intelligence in World War II", Journal of the American Statistical Association **42**, 72–91 (1947).

serial number analysis for German tank production

the Allies collected serial numbers on the chassis, engines, gearboxes, and bogie wheels of samples of tanks by inspecting captured tanks and examining captured records[3].

Monthly production of tanks by Germany.

	estimates	
date	conventional intelligence	serial number analysis
Jun. 1940	1000	169
Jun. 1941	1550	244
Aug. 1942	1550	327

[3] R. Ruggles and H. Brodie, "An empirical approach to economic intelligence in World War II", Journal of the American Statistical Association **42**, 72–91 (1947).

serial number analysis for German tank production

the Allies collected serial numbers on the chassis, engines, gearboxes, and bogie wheels of samples of tanks by inspecting captured tanks and examining captured records[3].

Monthly production of tanks by Germany.

	estimates		
date	conventional intelligence	serial number analysis	German records
Jun. 1940	1000	169	122
Jun. 1941	1550	244	271
Aug. 1942	1550	327	342

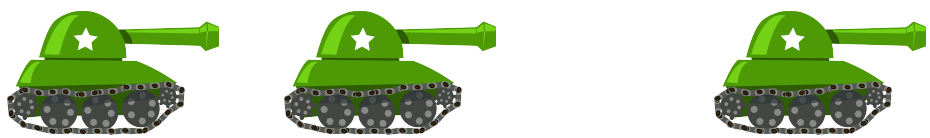
[3] R. Ruggles and H. Brodie, "An empirical approach to economic intelligence in World War II", Journal of the American Statistical Association **42**, 72–91 (1947).

the German tank problem

the German tank problem

problem statement
the German military has n tanks. each tank is inscribed with a unique serial number in $\{1, ..., n\}$.

as the Allies, we do not know n, but we captured (without replacement) a sample of k tanks with serial numbers $(s_1, ..., s_k)$.



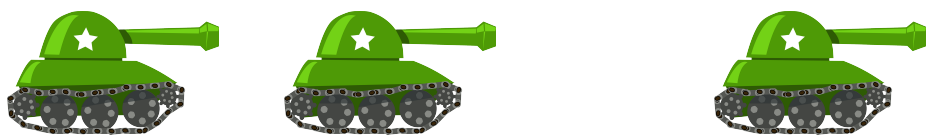
s_1 s_2 $\cdots$ s_k

objective: estimate n in consideration of the data $(s_1, ..., s_k)$.

the German tank problem

problem statement
the German military has n tanks. each tank is inscribed with a unique serial number in $\{1, ..., n\}$.

as the Allies, we do not know n, but we captured (without replacement) a sample of k tanks with serial numbers $(s_1, ..., s_k)$.

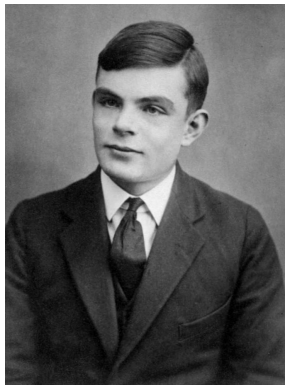


s_1 s_2 $\cdots$ s_k

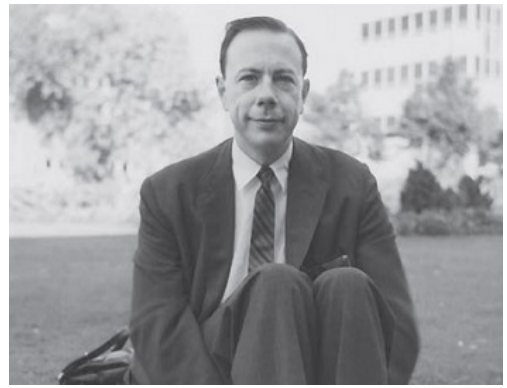
objective: estimate n in consideration of the data $(s_1, ..., s_k)$.

key assumption: all tanks in the population were equally likely to be captured.

the German tank problem



(a) Alan Turing



(b) Andrew Gleason

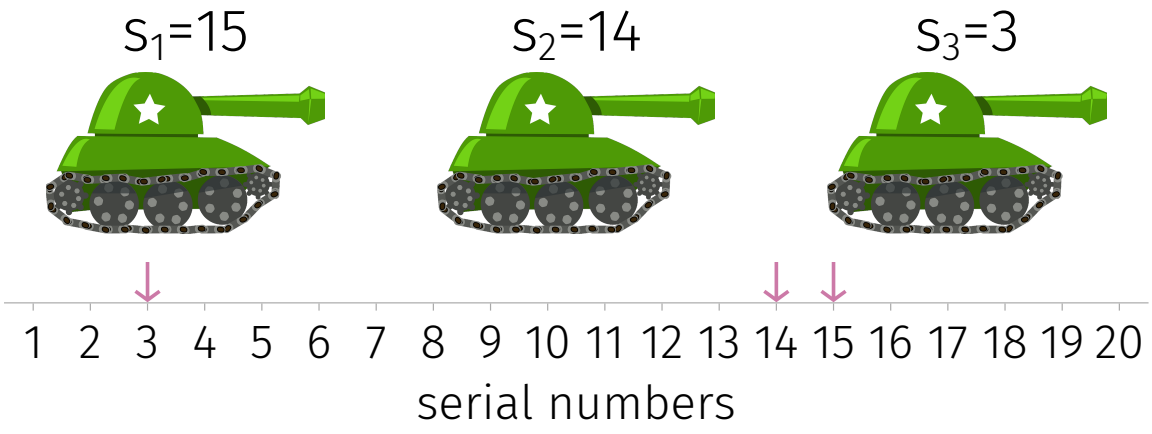
origins?

In 1942, Alan Turing and Andrew Gleason discussed a variant of the German tank problem, “how to best to estimate the total number of taxicabs in a town, having seen a random selection of their license numbers”, in a crowded restaurant in Washington DC[4,5].

[4] A. Hodges, “Alan Turing: the enigma”, in *Alan turing: the enigma* (Princeton University Press, 2014).

[5] M. Hall, “Alan Turing, Marshall Hall, and the Alignment of WW2 Japanese Naval Intercepts”, Notices of the AMS **61** (2014).

example



the data (s_1, s_2, s_3).

what is your estimate of n?

quantifying uncertainty in the tank population size

any estimate of the tank population size n from the data $(s_1, ..., s_k)$ is subject to uncertainty.

quantifying uncertainty in the tank population size

any estimate of the tank population size n from the data $(s_1, ..., s_k)$ is subject to uncertainty.

quantifying uncertainty in our estimate of n is important because high-stakes military decisions may be made on its basis.

🦕: ”I estimate n to be between 15 and 25 with 80% confidence.”

a Bayesian treatment of the German tank problem

this talk: the Bayesian approach to the German tank problem

- solution quantifies uncertainty by assigning a probability to each tank population size

a Bayesian treatment of the German tank problem

this talk: the Bayesian approach to the German tank problem

- solution quantifies uncertainty by assigning a probability to each tank population size
- provides an opportunity to incorporate prior information and/or beliefs about the tank population size into the solution.

prior work on the German tank problem

references

the frequentist approach. Goodman[6], and Clark, Gonye, and Miller[7].

for pedagogy. Champkin,[8], Johnson[9], Scheaffer, Watkins, Gnanadesikan, and Witmer[10], Berg[11].

the Bayesian approach. Roberts[12], Höhle and Held[13], and Linden, Dose, and Toussaint[14], Cocco, Monasson, and Zamponi[15], and Andrews[16].

---

[6] L. A. Goodman, "Serial number analysis", Journal of the American Statistical Association **47**, 622–634 (1952), L. A. Goodman, "Some practical techniques in serial number analysis", Journal of the American Statistical Association **49**, 97–112 (1954).

[7] G. Clark et al., "Lessons from the german tank problem", The Mathematical Intelligencer **43**, 19–28 (2021).

[8] C. G. Grajalez et al., "Great moments in statistics", Significance **10**, 21–28 (2013).

[9] R. W. Johnson, "Estimating the size of a population", Teaching Statistics **16**, 50–52 (1994).

[10] R. L. Scheaffer et al., *Activity-based statistics: student guide*, (Springer Science & Business Media, 2013).

[11] A. Berg, "Bayesian modeling competitions for the classroom", Revista Colombiana de Estadística **44**, 243–252 (2021).

[12] H. V. Roberts, "Informative stopping rules and inferences about population size", Journal of the American Statistical Association **62**, 763–775 (1967).

[13] M. Höhle and L. Held, *Bayesian estimation of the size of a population*, tech. rep. 499 (LMU Munich, Discussion Paper, 2006).

[14] W. Von der Linden et al., *Bayesian probability theory: applications in the physical sciences*, (Cambridge University Press, 2014).

[15] S. Cocco et al., *From statistical physics to data-driven modelling: with applications to quantitative biology*, (Oxford University Press, 2022).

[16] M. Andrews, *German tank problem: a bayesian analysis*, `https://www.mjandrews.org/blog/germantank`, Accessed: 2022-12-03.

a Bayesian treatment of the German tank problem

a Bayesian treatment of the German tank problem

modeling uncertainty

treat N as a discrete random variable

a probability mass function (PMF) of N assigns a probability to each possible tank population size $n \in \{0, 1, \ldots\}$.

[17] J. K. Ghosh et al., *An introduction to bayesian analysis: theory and methods*, Vol. 725 (Springer, 2006).

treat N as a discrete random variable

a probability mass function (PMF) of N assigns a probability to each possible tank population size $n \in \{0, 1, \ldots\}$.

💡 this probability is a measure of our degree of belief, perhaps with some basis in knowledge/data, that the tank population size is n.[17]

[17] J. K. Ghosh et al., *An introduction to bayesian analysis: theory and methods*, Vol. 725 (Springer, 2006).

treat N as a discrete random variable

a probability mass function (PMF) of N assigns a probability to each possible tank population size $n \in \{0, 1, ...\}$.

💡 this probability is a measure of our degree of belief, perhaps with some basis in knowledge/data, that the tank population size is n.[17]

the observed serial numbers $(s_1, ..., s_k)$ provide information about the tank population size. consequently, N has a:

- prior PMF
- posterior PMF

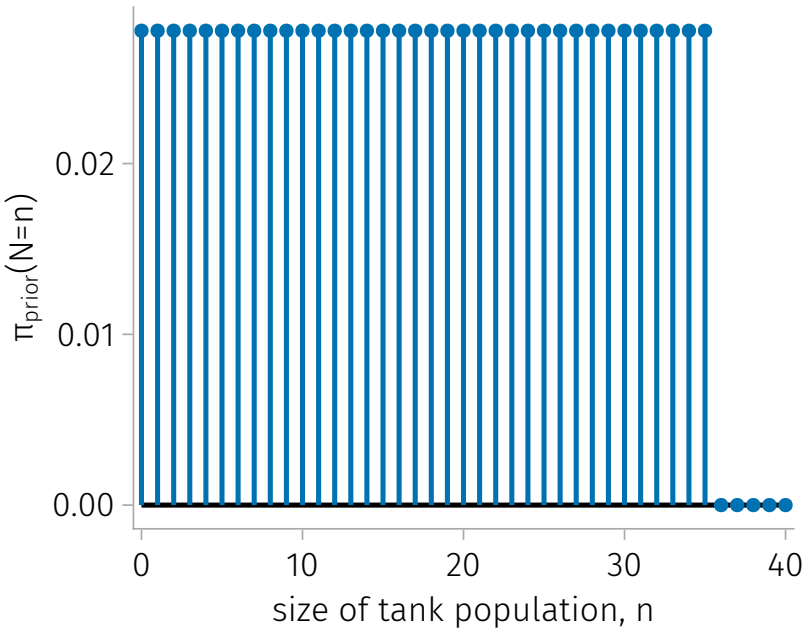
[17] J. K. Ghosh et al., *An introduction to bayesian analysis: theory and methods*, Vol. 725 (Springer, 2006).

a Bayesian treatment of the German tank problem

ingredient 1: the prior PMF

the prior PMF of N, $\pi_{\text{prior}}(N = n)$

expresses a combination of our subjective beliefs and objective knowledge about the total number of tanks N before the data $(s_1, \ldots, s_k)$ are collected and considered.

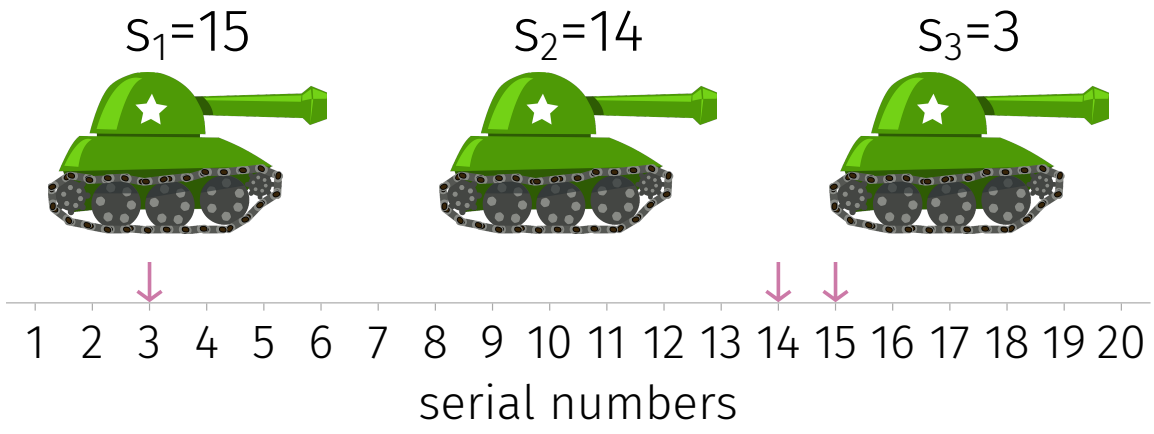


$$\pi_{\text{prior}}(N = n) = \frac{1}{n_{\text{max}} + 1} \mathcal{I}_{\{0,\ldots,n_{\text{max}}\}}(n).$$

a Bayesian treatment of the German tank problem

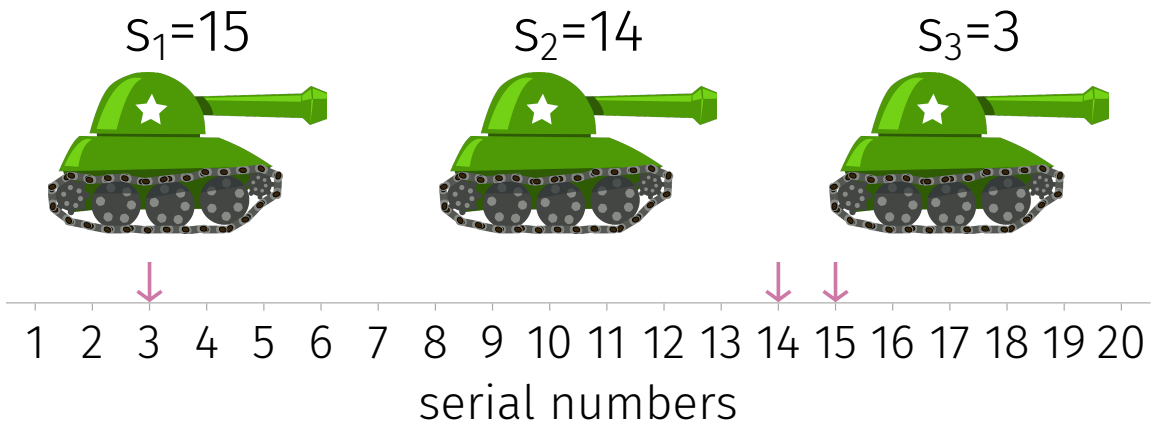
ingredient 2: the data

the data, $s^{(k)} := (s_1, \ldots, s_k)$



the data $s^{(3)} = (15, 14, 3)$.

the data, $s^{(k)} := (s_1, \ldots, s_k)$



the data $s^{(3)} = (15, 14, 3)$.

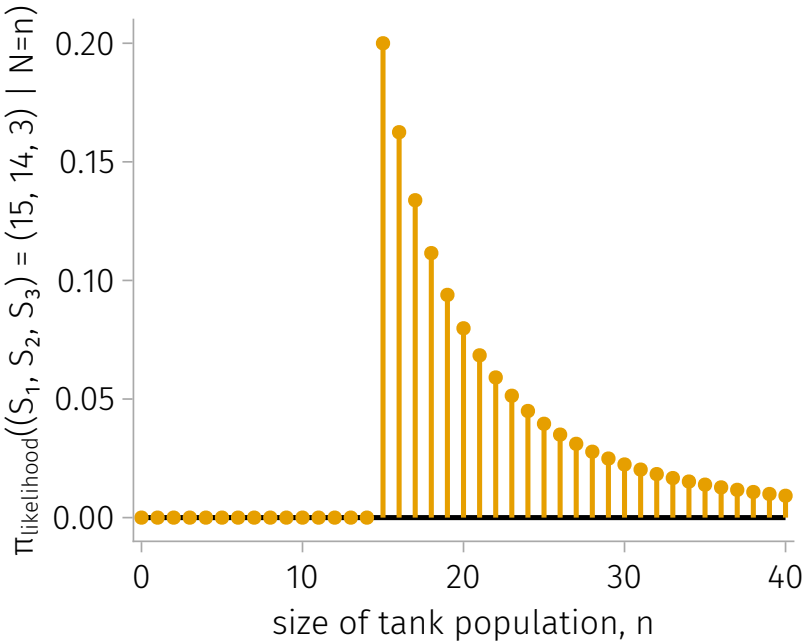
💡 $s^{(k)}$ is a realization of the discrete random vector $S^{(k)} := (S_1, \ldots, S_k)$.

a Bayesian treatment of the German tank problem

ingredient 3: the likelihood function

the likelihood function, $\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n)$

- specifies the probability of the data $S^{(k)} = s^{(k)}$ given the tank population size $N = n$
- quantifies the support the data $s^{(k)}$ lend for each tank population size n
- constructed from a probabilistic model of the data-generating process



the likelihood function, $\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n)$

a probabilistic model of the data-generating process: sequential capture of k tanks from a population of n tanks, without replacement.

the likelihood function, $\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n)$

a probabilistic model of the data-generating process: sequential capture of k tanks from a population of n tanks, without replacement.

perspective 1: a uniform distribution over outcomes
each outcome $s^{(k)}$ belonging to the sample space

$$\Omega_n^{(k)} := \{(s_1, ..., s_k)_{\neq} : s_i \in \{1, ..., n\} \text{ for all } i \in \{1, ..., k\}\}$$

is equally likely.

the likelihood function, $\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n)$

a probabilistic model of the data-generating process: sequential capture of k tanks from a population of n tanks, without replacement.

perspective 1: a uniform distribution over outcomes
each outcome $s^{(k)}$ belonging to the sample space

$$\Omega_n^{(k)} := \{(s_1, ..., s_k)_{\neq} : s_i \in \{1, ..., n\} \text{ for all } i \in \{1, ..., k\}\}$$

is equally likely.

the number of possible outcomes is

$$|\Omega_n^{(k)}| = n(n-1)\cdots(n-(k-1)) = n!/(n-k)! =: (n)_k$$

the likelihood function, $\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n)$

a probabilistic model of the data-generating process: sequential capture of k tanks from a population of n tanks, without replacement.

perspective 1: a uniform distribution over outcomes
each outcome $s^{(k)}$ belonging to the sample space

$$\Omega_n^{(k)} := \{(s_1, ..., s_k)_{\neq} : s_i \in \{1, ..., n\} \text{ for all } i \in \{1, ..., k\}\}$$

is equally likely.

the number of possible outcomes is

$$|\Omega_n^{(k)}| = n(n-1)\cdots(n-(k-1)) = n!/(n-k)! =: (n)_k$$

the DGP is then the uniform distribution

$$\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n) = \frac{1}{(n)_k} \mathcal{I}_{\Omega_n^{(k)}}\left(s^{(k)}\right).$$

the likelihood function, $\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n)$

a probabilistic model of the data-generating process: sequential capture of k tanks from a population of n tanks, without replacement.

perspective 2: a sequence of events, $S_1 = s_1, \ldots, S_k = s_k$

the likelihood function, $\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n)$

a probabilistic model of the data-generating process: sequential capture of k tanks from a population of n tanks, without replacement.

perspective 2: a sequence of events, $S_1 = s_1, \ldots, S_k = s_k$

$$\pi(S_i = s_i \mid N = n, S_1 = s_1, \ldots, S_{i-1} = s_{i-1}) = \frac{1}{n - (i-1)} \mathcal{I}_{\{1,\ldots,n\} \setminus \{s_1,\ldots,s_{i-1}\}}(s_i)$$

the likelihood function, $\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n)$

a probabilistic model of the data-generating process: sequential capture of k tanks from a population of n tanks, without replacement.

perspective 2: a sequence of events, $S_1 = s_1, \ldots, S_k = s_k$

$$\pi(S_i = s_i \mid N = n, S_1 = s_1, \ldots, S_{i-1} = s_{i-1}) = \frac{1}{n - (i-1)} \mathcal{I}_{\{1,\ldots,n\} \setminus \{s_1,\ldots,s_{i-1}\}}(s_i)$$

chain rule:

$$\pi(S_1 = s_1, \ldots, S_k = s_k \mid N = n) = \prod_{i=1}^{k} \pi(S_i = s_i \mid N = n, S_1 = s_1, \ldots, S_{i-1} = s_{i-1}).$$

the likelihood function, $\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n)$

a probabilistic model of the data-generating process: sequential capture of k tanks from a population of n tanks, without replacement.

perspective 2: a sequence of events, $S_1 = s_1, \ldots, S_k = s_k$

$$\pi(S_i = s_i \mid N = n, S_1 = s_1, \ldots, S_{i-1} = s_{i-1}) = \frac{1}{n - (i-1)} \mathcal{I}_{\{1,\ldots,n\} \setminus \{s_1,\ldots,s_{i-1}\}}(s_i)$$

chain rule:

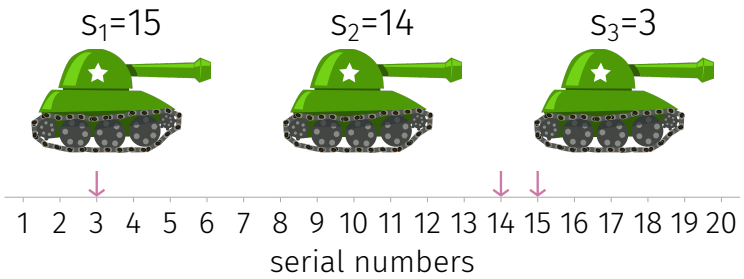
$$\pi(S_1 = s_1, \ldots, S_k = s_k \mid N = n) = \prod_{i=1}^{k} \pi(S_i = s_i \mid N = n, S_1 = s_1, \ldots, S_{i-1} = s_{i-1}).$$

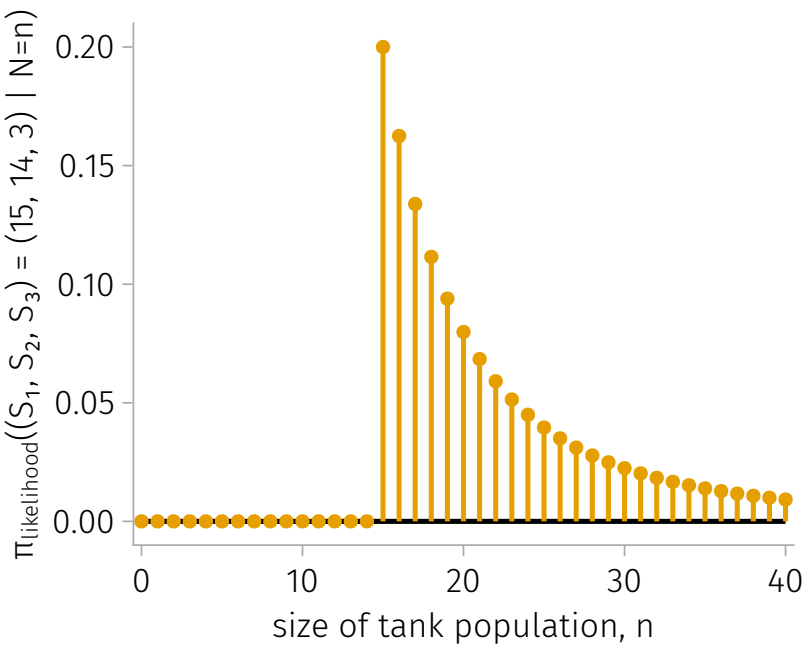
ie.

$$\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n) = \frac{1}{(n)_k} \mathcal{I}_{\Omega_n^{(k)}}\left(s^{(k)}\right).$$

the likelihood function, $\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n)$

$$\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n) = \frac{1}{(n)_k} \mathcal{I}_{\Omega_n^{(k)}}\left(s^{(k)}\right)$$





a Bayesian treatment of the German tank problem

turning the Bayesian crank to obtain the posterior PMF

the posterior PMF of N, $\pi_{\text{posterior}}(N = n \mid S^{(k)} = s^{(k)})$

💡 assigns a probability to each possible tank population size n in consideration of its consistency with (1) the data $(s_1, \ldots, s_k)$, according to the likelihood, and (2) our prior beliefs/knowledge encoded in $\pi_{\text{prior}}(N = n)$.

the posterior PMF of N, $\pi_{\text{posterior}}(N = n \mid S^{(k)} = s^{(k)})$

💡 assigns a probability to each possible tank population size n in consideration of its consistency with (1) the data $(s_1, \ldots, s_k)$, according to the likelihood, and (2) our prior beliefs/knowledge encoded in $\pi_{\text{prior}}(N = n)$.

Bayes' theorem:

$$\pi_{\text{posterior}}(N = n \mid S^{(k)} = s^{(k)}) = \frac{\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n)\pi_{\text{prior}}(N = n)}{\pi_{\text{data}}(S^{(k)} = s^{(k)})},$$

with

$$\pi_{\text{data}}(S^{(k)} = s^{(k)}) = \sum_{n'=0}^{\infty} \pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n')\pi_{\text{prior}}(N = n').$$

the posterior PMF of N, $\pi_{\text{posterior}}(N = n \mid S^{(k)} = s^{(k)})$

💡 assigns a probability to each possible tank population size n in consideration of its consistency with (1) the data $(s_1, \ldots, s_k)$, according to the likelihood, and (2) our prior beliefs/knowledge encoded in $\pi_{\text{prior}}(N = n)$.

Bayes' theorem:

$$\pi_{\text{posterior}}(N = n \mid S^{(k)} = s^{(k)}) = \frac{\pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n)\pi_{\text{prior}}(N = n)}{\pi_{\text{data}}(S^{(k)} = s^{(k)})},$$

with

$$\pi_{\text{data}}(S^{(k)} = s^{(k)}) = \sum_{n'=0}^{\infty} \pi_{\text{likelihood}}(S^{(k)} = s^{(k)} \mid N = n')\pi_{\text{prior}}(N = n').$$

$\pi_{\text{posterior}}(N = n \mid S^{(k)} = s^{(k)})$ is the raw solution to the German tank problem!

the posterior PMF of N, $\pi_{\text{posterior}}(N = n \mid S^{(k)} = s^{(k)})$

let

$$m^{(k)} = \max_{i \in \{1,\ldots,k\}} s_i.$$

the posterior PMF of N, $\pi_{\text{posterior}}(N = n \mid S^{(k)} = s^{(k)})$

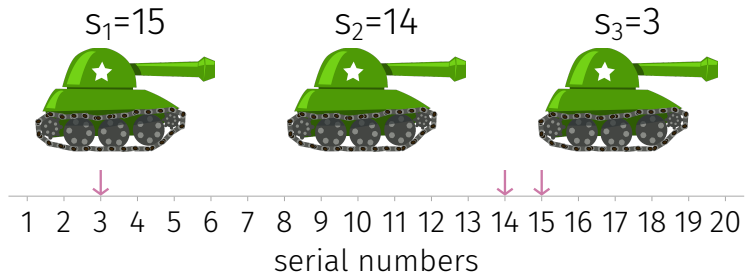
let

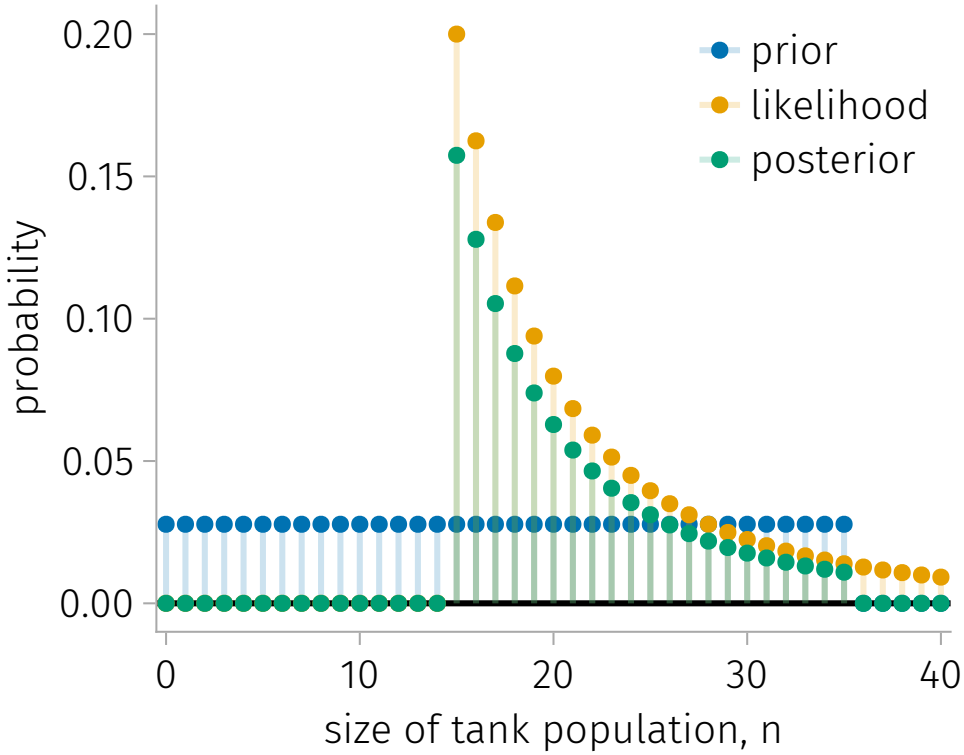
$$m^{(k)} = \max_{i \in \{1,\ldots,k\}} s_i.$$

$$\begin{aligned}\pi_{\text{posterior}}(N = n \mid S^{(k)} = s^{(k)}) &= \frac{(n)_k^{-1} \pi_{\text{prior}}(N = n)}{\sum_{n'=m^{(k)}}^{\infty} (n')_k^{-1} \pi_{\text{prior}}(N = n')} \mathcal{I}_{\{m^{(k)}, m^{(k)}+1, \ldots\}}(n) \\ &= \pi_{\text{posterior}}(N = n \mid M^{(k)} = m^{(k)})\end{aligned}$$

ie. the data $s^{(k)}$ can be distilled to $m^{(k)}$.

the posterior PMF of N, $\pi_{\text{posterior}}(N = n \mid S^{(k)} = s^{(k)})$





summarizing posterior PMF of N

the α-high-mass subset[18]

$$\mathcal{H}_\alpha := \{n' : \pi_{\text{posterior}}(N = n' \mid M^{(k)} = m^{(k)}) \geq \pi_\alpha\} \tag{1}$$

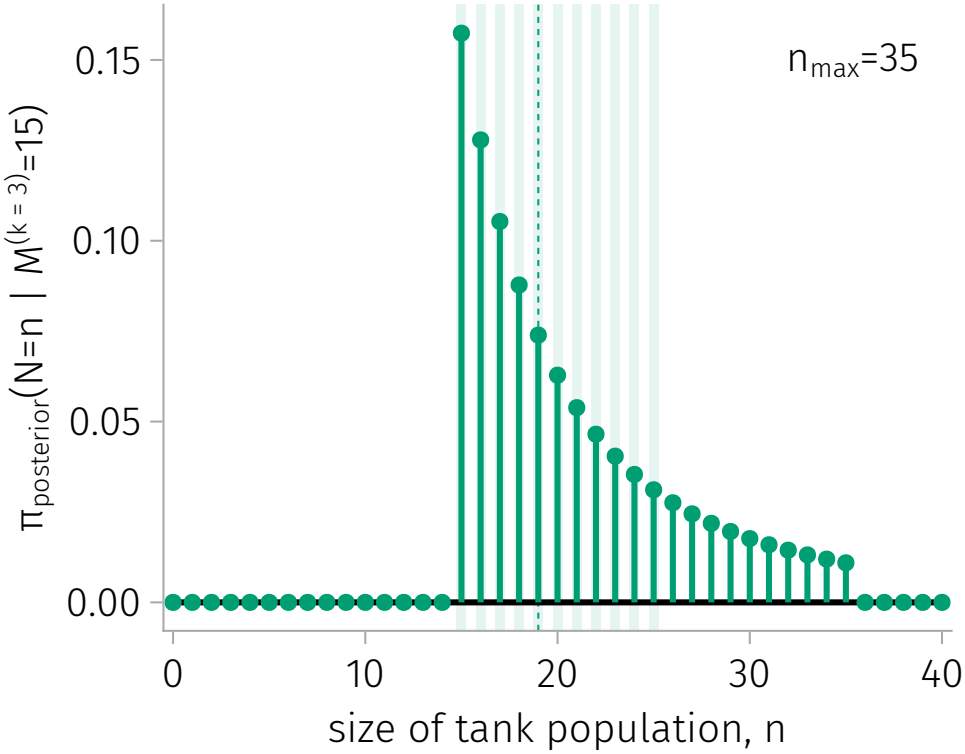
where π_α is the largest mass to satisfy

$$\pi_{\text{posterior}}(N \in \mathcal{H}_\alpha \mid M^{(k)} = m^{(k)}) \geq 1 - \alpha. \tag{2}$$

$\mathcal{H}_\alpha$ is the smallest subset of $\{0, ...\}$ to (i) contain at least a fraction $1 - \alpha$ of the posterior mass of N and (ii) ensure every tank population size belonging to it is more probable than any outside of it.

[18] R. J. Hyndman, "Computing and graphing highest density regions", The American Statistician **50**, 120–126 (1996).

the posterior PMF of N, $\pi_{\text{posterior}}(N = n \mid S^{(k)} = s^{(k)})$



median: 19

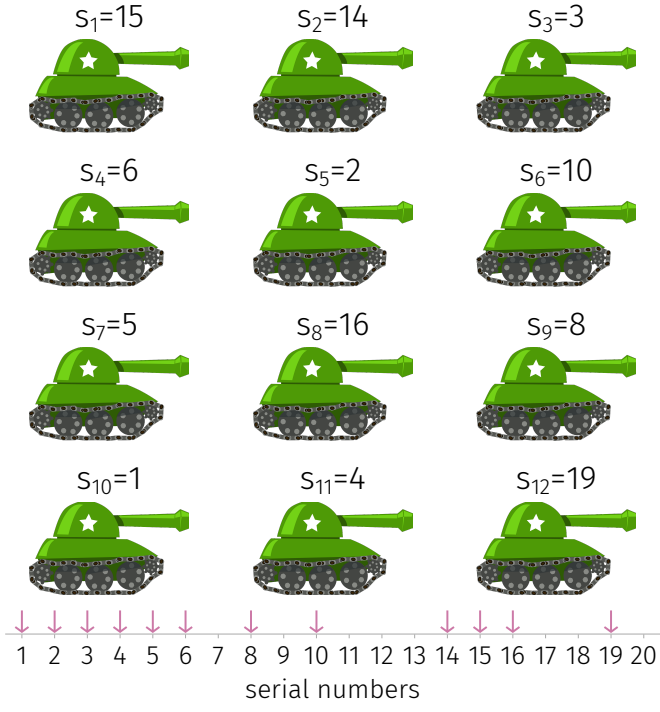
high-mass credible subset $\mathcal{H}_{0.2} = \{15, \ldots, 25\}$

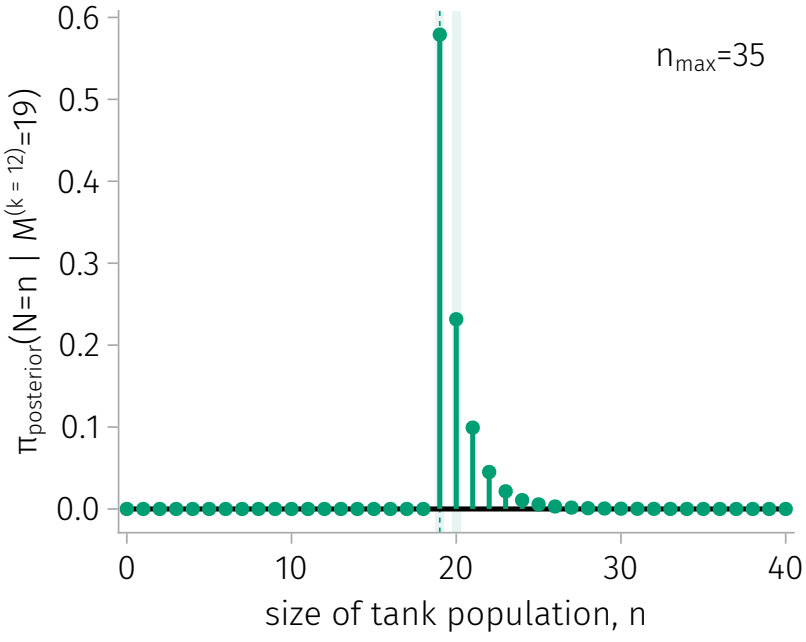
eg. $\pi_{\text{posterior}}(N > 30 \mid M^{(3)} = 15) \approx 0.066$.

a Bayesian treatment of the German tank problem

updating our posterior PMF with more data

what if we capture more tanks?

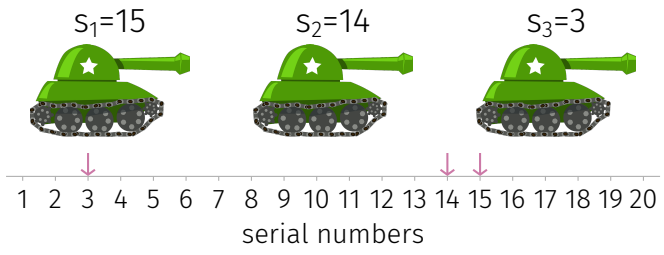


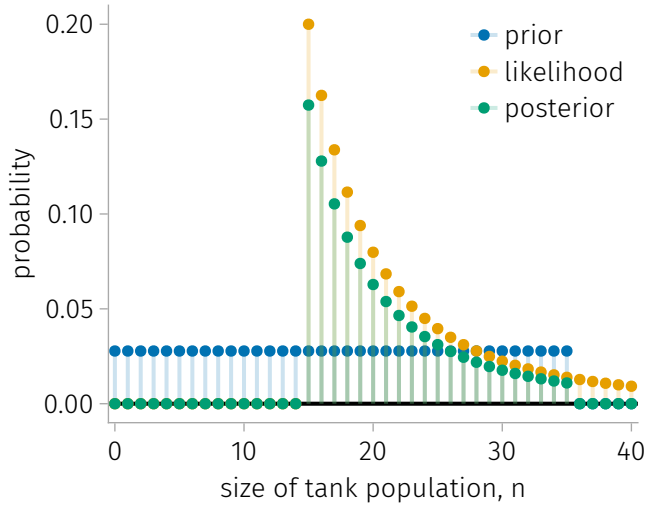


conclusions

a Bayesian treatment of the German tank problem

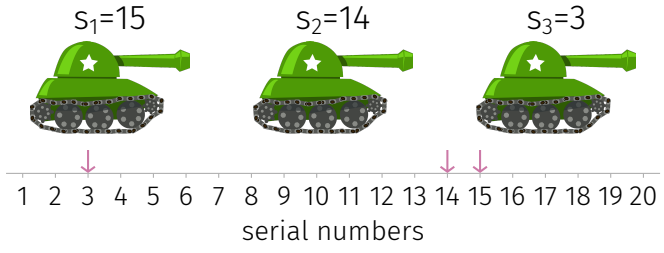
incorporate prior assumptions, quantify uncertainty about the tank population size, *N*

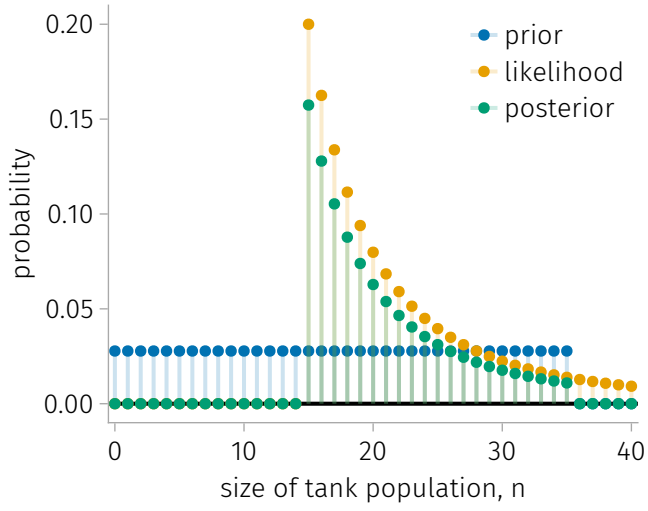




a Bayesian treatment of the German tank problem

incorporate prior assumptions, quantify uncertainty about the tank population size, *N*





limitation: uncertainty about tank-capturing model neglected. selection bias may be present.

the German tank problem in other contexts

estimating the size of some finite, “hidden” set, eg. the:

- number of taxicabs in a city
- number of accounts at a bank
- number of aircraft operations at an airport
- extent of leaked classified government communications
- time-coverage of historical records of extreme events like floods
- number of iPhones produced by Apple
- length of a short-tandem repeat allele
- size of a social network
- number of cases in court
- lifetime of a flower of a plant
- duration of existence of a species

tangentially related: mark and recapture methods in ecology to estimate the size of an animal population.